Statistica Neerlandica (2002) Vol. 56, nr. 2, pp. 233–242

Nonparametric estimation of a discontinuity in regression

Astrid Dempfle[1] and Winfried Stute[2]

Mathematical Institute, University of Giessen, Arndtstrasse 2, D-35392 Giessen, Germany

We propose and study a new method to nonparametrically estimate a discontinuity of a regression function. The optimal rate of convergence n^{-1} is obtained under minimal assumptions. No smoothing is required.

Key Words and Phrases: changepoint, regression, rate of convergence.

1 Introduction and main result

In this paper we propose and study a new method to estimate nonparametrically a possible discontinuity of a regression function

$$m(x) = \mathbb{E}(Y|X = x),$$

where (X, Y) is a bivariate random vector with joint (unknown) distribution H and Y has a finite expectation. To know such jumps is of some practical importance, since they indicate that there will be an abrupt change in the expected output Y once the independent variable X approaches a certain threshold, say θ. Apparently HINKLEY (1969a, b) was among the first to study, in a parametric framework, this problem, for fixed design and normal errors. He showed that it is possible to estimate θ within the order n^{-1}, where n is sample size. HALL and TITTERINGTON (1992) and MÜLLER (1992) initiated the estimation of a changepoint in regression in a completely non-parametric framework. Assuming that m has a jump at θ but is (very) smooth otherwise, their approach is based on a comparison, at each x, of one-sided nonparametric estimators of $m(x-)$ and $m(x+)$, the left- and right-hand limits of m at x. The estimator of θ is then obtained as the maximizer of an appropriate (data based) discrepancy measuring the 'roughness' of m at each x. MÜLLER (1992) obtained, again for fixed (equidistant) design and independent homoscedastic errors, that under some heavy smoothness assumptions θ may be estimated within the order $(n^{-1+\delta})$. For normal errors and under the assumption that θ is among the design variables, LOADER (1996) refined MÜLLER's (1992) smoothing methodology and obtained the optimal rate n^{-1}. MÜLLER and SONG (1997) got the same rate through a two-step procedure, under less restrictive assumptions than LOADER (1996). Modifications and

[1] demple@med.uni-marburg.de
[2] winfried.stute@math.uni-giessen.de

extensions of this approach can be found in GIJBELS *et al.* (1999), JOSE and ISMAIL (1999), KOCH and POPE (1997), LIU *et al.* (1997), QIU and YANDELL (1998) and WU and CHU (1993), among others. We finally mention KHMALADZE *et al.* (1997), who considered, for a multivariate X, estimating a change set G such that the conditional distribution of Y given $X = x$ equals a given P_1 when $x \notin G$ and P_2 when $x \in G$. Their method consists of maximizing a likelihood function parametrized by a Vapnik–Chervonenkis class, under proper regularity and boundedness assumptions on the underlying density.

It is worthwhile summarizing the assumptions and key steps needed for the 'smoothing' approach:

1. m should be (very) smooth outside of θ.
2. Smoothing technology requires estimation of the two one-sided limits $m(x-)$ and $m(x+)$, at each x.
3. In the fixed design case, the x's are distributed regularly with known support. For random design, the X's need to satisfy additional conditions, e.g., as to the support or the tails.
4. Typically, the errors are independent of the X's and homoscedastic.
5. Smoothing parameters to be chosen depend on other (unknown) model quantities.

In this paper, only the case of stochastic design will be discussed in detail, but fixed design may be dealt with in a similar way. Now, assuming that Y has a finite first moment, we have the decomposition

$$Y = m(X) + \varepsilon, \qquad \text{where } \mathbb{E}(\varepsilon | X) = 0.$$

Already the simplest example of a count variable, namely a 0-1 binary variable Y, leading to

$$m(x) = \mathbb{P}(Y = 1 | X = x),$$

shows that the conditional variance of ε given x equals $m(x)(1 - m(x))$ and hence depends on x. In other words, though ε is orthogonal to X, any independence assumption on ε and X will unavoidably limit the applicability of a proposed method. Therefore, any method which claims to work in a reasonably broad setup, should not use the ε's at all. Also assumptions on the tails or the support of the X's should be avoided if one wants to exhibit m locally.

In this paper we propose a method which avoids all this – it is a one-step procedure, computationally simple, which circumvents smoothing and yields the optimal rate n^{-1} under minimal assumptions on the model. It constitutes a modification of an approach studied in FERGER and STUTE (1992) to detect a change in distribution of a sequence of independent random variables. It has its origin in U-statistics rather than smoothing methodology. See also DÜMBGEN (1991). Our main emphasis will be on a very simple model for m, namely

$$m(x) = a1_{\{x \leqslant \theta\}} + b1_{\{x > \theta\}}. \tag{1}$$

Here, the levels a and b are unknown, as is the (marginal) distribution F of X. Clearly, m is discontinuous at θ if and only if $a \neq b$. This model seems to be very restrictive at first sight, but as we will see later the results obtained under (1) will be useful in quite a general setting. Our estimate of θ will be computed from n independent replications (X_i, Y_i), $1 \leqslant i \leqslant n$, of (X, Y). A necessary assumption to be made is

$$0 < F(\theta) < 1, \tag{2}$$

since, e.g., in the case $F(\theta) = 0$, we don't observe any data left to θ so that a detection of θ will be impossible. The main result of our paper asserts that our estimator of θ, say θ_n, satisfies

$$n(\theta_n - \theta) = O_{\mathbb{P}}(1). \tag{3}$$

We already mentioned that our approach circumvents smoothing techniques. In our view, a discontinuity of a regression function provides much more information through data than in the smooth case. This is expressed through the rate n^{-1} in (3). Estimation of θ via a smoothing may blur the relevant information resulting in worse rates, unless one assumes heavy regularity assumptions.

We now present a brief idea of our approach. Along with m consider the function

$$m_U(t) = a1_{\{t \leqslant F(\theta)\}} + b1_{\{t > F(\theta)\}}, \qquad 0 < t < 1.$$

The function m_U is the regression function corresponding to the pair $(F(X), Y)$. Note that $F(X)$ is uniformly distributed on (0, 1) whenever F is continuous. We will introduce a function r defined on (0, 1) such that $\mu = F(\theta)$ is the maximizer (or minimizer) of r. This approach works for general (X, Y), but r takes on a particular and insightful form when (1) holds. It will turn out that r allows for an empirical substitute, say r_n. Its maximizer μ_n will finally lead to an estimator of θ:

$$\theta_n = F_n^{-1}(\mu_n),$$

with F_n^{-1} denoting the quantile function pertaining to the empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leqslant x\}}$$

of the X_i's.

To define r, recall H, the joint distribution of (X, Y). Put, for $0 < t < 1$,

$$r(t) = \int\!\!\int (y_1 - y_2) 1_{\{x_1 \leqslant F^{-1}(t), x_2 > F^{-1}(t)\}} H(dx_1, dy_1) H(dx_2, dy_2).$$

The idea behind r is that for each fixed t we compare the mean values of Y subject to X being less than or equal to or larger than the t quantile of F. For a continuous F, we obtain

$$\begin{aligned}
r(t) &= \int\int (m(x_1) - m(x_2)) 1_{\{x_1 \leqslant F^{-1}(t), x_2 > F^{-1}(t)\}} F(dx_1) F(dx_2) \\
&= \int_0^1 \int_0^1 (m_U(z_1) - m_U(z_2)) 1_{\{z_1 \leqslant t, z_2 > t\}} dz_1 dz_2 \\
&= \int_t^1 \int_0^t [m_U(z_1) - m_U(z_2)] dz_1 dz_2 \\
&= \int_0^t m_U(z) dz - t \int_0^1 m_U(z) dz \\
&= (a - b)[-tF(\theta) + t 1_{\{t \leqslant F(\theta)\}} + F(\theta) 1_{\{t > F(\theta)\}}]. \qquad (4)
\end{aligned}$$

From the last expression we see that r has triangular form with

$$F(\theta) = \text{argmax}\, r(t)$$

provided that $a > b$. Otherwise, replace argmax by argmin. Note that r is constant if $a = b$, i.e., if there are no jumps. This observation may be useful also when one wants to test if there is a jump at all, the 'hypothesis of no jump' being rejected when the maximal deviation of r_n to be defined below exceeds a critical value.

Coming back to the estimation of θ, the empirical version of r is clearly given by

$$\begin{aligned}
r_n(t) &= \int\int (y_1 - y_2) 1_{\{x_1 \leqslant F_n^{-1}(t), x_2 > F_n^{-1}(t)\}} H_n(dx_1, dy_1) H_n(dx_2, dy_2) \\
&= n^{-2} \sum_{i=1}^n \sum_{j=1}^n (Y_i - Y_j) 1_{\{X_i \leqslant F_n^{-1}(t), X_j > F_n^{-1}(t)\}},
\end{aligned}$$

with H_n denoting the empirical distribution of (X_i, Y_i), $1 \leqslant i \leqslant n$. Introducing the order statistics $X_{1:n} \leqslant \ldots \leqslant X_{n:n}$ of the X-sample with pertaining concomitants $Y_{[1:n]}, \ldots, Y_{[n:n]}$, the last double sum may be rewritten, for $t \in \Delta_n = \{1/n, 2/n, \ldots, (n-1)/n\}$, as

$$r_n(t) = n^{-1} \left[\sum_{i=1}^{nt} Y_{[i:n]} - t \sum_{i=1}^n Y_i \right].$$

We see that r_n is the normalized tied-down partial sum process of the concomitants. Generally the concomitants are dependent random variables, so that its stochastic structure differs from the usual (tied down) partial sum process of independent random variables covered by Donsker's invariance principle. Rather we shall employ the facts, see Lemma 2.1 in STUTE and WANG (1993), that the concomitants are independent conditionally on the order statistics and that the conditional structure of $(X_{i:n}, Y_{[i:n]})$ is the same as of (X, Y). Along with r_n we also need the conditional expectation process of r_n:

$$\tilde{r}_n(t) = \mathbb{E}[r_n(t)|X_{1:n}, \ldots, X_{n:n}]$$
$$= \frac{1}{n}\left[\sum_{i=1}^{nt} \mathbb{E}(Y_{[i:n]}|X_{i:n}) - t\sum_{i=1}^{n} \mathbb{E}(Y_{[i:n]}|X_{i:n})\right],$$

from which we easily get

$$\tilde{r}_n(t) = (a-b)[-tF_n(\theta) + t1_{\{t\leqslant F_n(\theta)\}} + F_n(\theta)1_{\{t>F_n(\theta)\}}], \tag{5}$$

the empirical analogue of (4). Also $\tilde{r}_n$ has a triangular form with

$$F_n(\theta) = \operatorname{argmax} \tilde{r}_n$$

if $a > b$. The function $\tilde{r}_n$ plays some role in our approach since we need to center r_n conditionally on the X's in order to bring some martingale structure into play.

THEOREM. Assume that model (1) holds true, and that

(A1) the conditional variance of Y given $X = x$ is bounded:

$$\operatorname{Var}(Y|X = x) \leqslant \sigma^2 < \infty$$

(A2) the (marginal) distribution of X is continuously differentiable in a (small) neighbourhood of θ with $F'(\theta) > 0$.

Then we have

$$n(\theta_n - \theta) = O_{\mathbb{P}}(1).$$

Condition (A1) is trivially satisfied in the homoscedastic case whenever Y has a finite second moment. In the heteroscedastic case (A1) is close to being necessary also, since an unbounded $\sigma^2(x)$ may then cause extremely varying Y's so that it will be impossible to discriminate between two levels of m. (A2) is also close to being necessary, since it just guarantees that there are enough X-data both right and left of θ. Clearly, (2) is implied by (A2).

Note that no other conditions are required. In particular, the method works without estimating and comparing the 'levels' of m. Rather the idea behind our method is to exhibit any significant 'triangular structure' through r_n. Our Theorem shows that this works under (1). For further discussion assume $a > b$. It is easy to see that the Theorem also holds, when m is more general than (1), namely if

$$m(x) \geqslant a \text{ for } x \leqslant \theta \text{ and } m(x) \leqslant b \text{ for } x > \theta.$$

Among this class of m's, the model (1) is the most difficult one, since an increase (decrease) of $m(x)$ left (right) of θ only improves the feasibility to discriminate between two different levels of m.

What makes our approach even more appealing is the fact that it works also locally on subintervals, i.e., rather than looking at all data, we may modify our r_n to strategically investigate a neighborhood of any given empirical quantile. In other words, we may fix $s < u$ in Δ_n and base a search for a discontinuity of m in $[X_{sn:n}, X_{un:n}]$ on the process

$$r_n^0(t) = r_n^0(t;\, u,\, s) = \frac{1}{n}\left[\sum_{i=ns}^{nt} Y_{[i:n]} - \frac{t-s}{u-s}\sum_{i=ns}^{nu} Y_{[i:n]}\right],$$

where $s \leqslant t \leqslant u$. If s and u are such that the data form a fraction of the whole sample, the Theorem may be applied to this case as well, provided that (1) holds locally. This, however, holds true (approximately) if m has a discontinuity at some θ but is continuously differentiable in the right and left neighborhood of θ. In this sense, we agree with MÜLLER and SONG (1997), p. 324, that model (1) is the prototype of a jump regression model.

Our final comment is on the applicability of the Theorem. The distributional convergence of θ_n may be obtained from a detailed study of the rescaled process

$$R_n(t) = n\left[r_n\left(F_n\left(\theta + \frac{t}{n}\right)\right) - r_n(F_n(\theta))\right], \qquad t \in \mathbb{R}.$$

To get the distributional convergence of $n(\theta_n - \theta)$ towards a (non-degenerate) limit, one has to show that

(i) R_n weakly converges in the Skorokhod space $D[-M,\, M]$ for all finite M
(ii) $n(\theta_n - \theta)$ is stochastically bounded
(iii) the argmax (or argmin) functional is continuous.

We see that (ii), which is covered by our Theorem, is an essential part of this program. Actually, in (i), R_n can only be shown to converge on compacta and not on the compactified real line. Hence, (ii) will be needed to guarantee that R_n only needs to be studied on compacta. A full treatment of this issue will appear elsewhere, since it is beyond the scope of this paper.

2 Proof

PROOF OF THE THEOREM. We assume that $a > b$ so that μ is a maximizer of r. We need to show that for a given positive ε there exists some $T > 0$ such that for all large $n \geqslant 1$

$$\mathbb{P}(n|\theta_n - \theta| \geqslant T) \leqslant \varepsilon.$$

We deal only with the upper tails. For this, write

$$\{n(\theta_n - \theta) \geqslant T\} = \left\{F_n^{-1}(\mu_n) \geqslant \theta + \frac{T}{n}\right\}$$

$$\subset \left\{\mu_n \geqslant F_n\left(\theta + \frac{T}{n}\right)\right\} = \bigcup_{l = nF_n\left(\theta + \frac{T}{n}\right)}^{n} \left\{\mu_n = \frac{l}{n}\right\}.$$

Since μ_n is a maximizer of r_n, we get for each of the above l:

$$\left\{\mu_n = \frac{l}{n}\right\} \subset \left\{r_n\left(\frac{l}{n}\right) - r_n(F_n(\theta)) \geqslant 0\right\}$$
$$= \left\{K_l \geqslant n\left[\tilde{r}_n(F_n(\theta)) - \tilde{r}_n\left(\frac{l}{n}\right)\right]\right\},$$

with

$$K_l = n\left[r_n\left(\frac{l}{n}\right) - r_n(F_n(\theta)) - \tilde{r}_n\left(\frac{l}{n}\right) + \tilde{r}_n(F_n(\theta))\right].$$

Because of

$$l \geqslant nF_n\left(\theta + \frac{T}{n}\right) \geqslant nF_n(\theta)$$

we obtain from (5)

$$n\left[\tilde{r}_n(F_n(\theta)) - \tilde{r}_n\left(\frac{l}{n}\right)\right] = (a-b)F_n(\theta)(l - nF_n(\theta)) =: \gamma(l).$$

Since by assumption $a > b$ we have $\gamma(l) > 0$ on the set $\{F_n(\theta) > 0\}$. Moreover, the $\gamma(l)$ are nondecreasing in l. Finally, equation (5) yields the representation

$$K_l = \sum_{j=nF_n(\theta)+1}^{l} (Y_{[j:n]} - m(X_{j:n})) - \left(\frac{l}{n} - F_n(\theta)\right)\sum_{i=1}^{n}(Y_i - m(X_i)).$$

Summarizing we thus get

$$\mathbb{P}(n(\theta_n - \theta) \geqslant T) \leqslant \mathbb{P}(n(\theta_n - \theta) \geqslant T,\ F_n(\theta) > 0) + \mathbb{P}(F_n(\theta) = 0)$$
$$\leqslant \mathbb{P}\left(\max_l \gamma^{-1}(l) \sum_{j=nF_n(\theta)+1}^{l} (Y_{[j:n]} - m(X_{j:n})) \geqslant \frac{1}{2},\ F_n(\theta) > 0\right) \tag{6}$$
$$+ \mathbb{P}\left(\max_l \gamma^{-1}(l)\left(\frac{l}{n} - F_n(\theta)\right)\left|\sum_{i=1}^{n}(Y_i - m(X_i))\right| \geqslant \frac{1}{2},\ F_n(\theta) > 0\right) \tag{7}$$
$$+ \mathbb{P}(F_n(\theta) = 0). \tag{8}$$

Since $0 < F(\theta) < 1$ by assumption, the law of large numbers guarantees that the last probability tends to zero as $n \to \infty$. To bound (6), first condition on the X's. Lemma 2.1 from STUTE and WANG (1993) implies that the summands $Y_{[j:n]} - m(X_{j:n})$ are conditionally independent and centered.

Since the $\gamma(l)$ are measurable w.r.t. the X's and nondecreasing, we may apply the Hájek–Rényi inequality, in a conditional setup, to get

$$P\left(\max_l \gamma^{-1}(l) \sum_{j=nF_n(\theta)+1}^{l} (Y_{[j:n]} - m(X_{j:n})) \geqslant \frac{1}{2},\ F_n(\theta) > 0 \mid X_{1:n}, \ldots, X_{n:n}\right) \tag{9}$$

$$\leqslant \frac{4 \cdot 1_{\{F_n(\theta)>0\}}}{\gamma^2\left(nF_n\left(\theta+\frac{T}{n}\right)\right)} \operatorname{Var}\left(\sum_{j=nF_n(\theta)+1}^{nF_n(\theta+\frac{T}{n})} (Y_{[j:n]} - m(X_{j:n}))|X_{1:n}, \ldots, X_{n:n}\right) \tag{10}$$

$$+ 4 \cdot 1_{\{F_n(\theta)>0\}} \sum_{j=nF_n\left(\theta+\frac{T}{n}\right)+1}^{n} \gamma^{-2}(j)\operatorname{Var}(Y_{[j:n]}|X_{j:n}). \tag{11}$$

To bound (10) we may assume w.l.o.g. that

$$nF_n(\theta) < nF_n\left(\theta+\frac{T}{n}\right),$$

since otherwise the sum is empty. By assumption (A1),

$$\operatorname{Var}(Y_{[j:n]}|X_{j:n}) \leqslant \sigma^2.$$

Conclude that (10) is less than or equal to

$$\frac{4\sigma^2 1_{\{F_n(\theta)>0\}}}{(a-b)^2 F_n^2(\theta)\left[nF_n\left(\theta+\frac{T}{n}\right) - nF_n(\theta)\right]}. \tag{12}$$

Put

$$c_T = nF_n\left(\theta+\frac{T}{n}\right) - nF_n(\theta),$$

a binomial random variable with parameter

$$p_n = F\left(\theta+\frac{T}{n}\right) - F(\theta).$$

The integral of (12) is less than or equal to

$$(a-b)^{-2}\int_{\{nF_n(\theta)\geqslant 1, c_T\geqslant 1\}} \frac{d\mathbb{P}}{F_n^2(\theta)c_T}$$

$$\leqslant \frac{n^2}{(a-b)^2}\sqrt{\int_{\{nF_n(\theta)\geqslant 1\}} \frac{d\mathbb{P}}{(nF_n(\theta))^4}}\sqrt{\int_{\{c_T\geqslant 1\}} c_T^{-2}\,d\mathbb{P}},$$

by Cauchy–Schwarz. Now use the simple bound

$$k^{-2} \leqslant \frac{6}{(k+1)(k+2)}$$

to get

$$\int_{\{c_T \geqslant 1\}} c_T^{-2} d\mathbb{P} = \sum_{k=1}^{n} k^{-2} \binom{n}{k} p_n^k (1 - p_n)^{n-k} \leqslant \frac{6}{(n+2)(n+1)p_n^2}.$$

By (A2) and the mean value theorem, the last term can be made arbitrarily small if we choose T large enough. Similarly, use the trivial bound

$$k^{-4} \leqslant \frac{128}{(k+1)(k+2)(k+3)(k+4)}$$

to get

$$n^4 \int_{\{nF_n(\theta) \geqslant 1\}} \frac{d\mathbb{P}}{(nF_n(\theta))^4} \leqslant \frac{128n^4}{(n+4)(n+3)(n+2)(n+1)F^4(\theta)} \leqslant \frac{128}{F^4(\theta)}.$$

Altogether this shows that the integral of (12) can be made arbitrarily small by letting $T \to \infty$.

We now bound (11). From the conditional independence we obtain that (11) is less than or equal to

$$4\sigma^2 1_{\{F_n(\theta)>0\}} \sum_{j=nF_n\left(\theta+\frac{T}{n}\right)+1}^{n} \gamma^{-2}(j) \leqslant 1_{\{F_n(\theta)>0\}} \frac{4\sigma^2}{(a-b)^2 F_n^2(\theta)} \sum_{k=k_0}^{\infty} k^{-2},$$

where $k_0 \equiv nF_n(\theta + (T/n)) - nF_n(\theta) + 1$.

Now use

$$\sum_{k=k_0}^{\infty} k^{-2} \leqslant \sum_{k=k_0}^{\infty} \frac{2}{k(k+1)} = \frac{2}{k_0} = \frac{2}{c_T + 1}$$

to finally get that the expectation of (11) is less than or equal to

$$4\sigma^2 (a-b)^{-2} \int_{\{F_n(\theta)>0\}} \frac{2d\mathbb{P}}{F_n^2(\theta)(c_T+1)}$$

which as already has been shown can be made arbitrarily small by letting $T \to \infty$. Summarizing this shows that (6) can be made small for large enough T. The proof of the Theorem will be completed by bounding (7).

By definition of $\gamma(l)$, the maximum equals

$$\frac{1}{n(a-b)F_n(\theta)} \left| \sum_{i=1}^{n} (Y_i - m(X_i)) \right|$$

which, however, goes to zero by the law of large numbers.

The proof is complete □

References

DÜMBGEN, L. (1991), The asymptotic behavior of some nonparametric changepoint estimators, *Annals of Statistics* **19**, 1471–1495.

FERGER, D. and W. STUTE (1992), Convergence of changepoint estimators, *Stochastic Processes and their Applications* **42**, 345–351.

GIJBELS, I., P. HALL and A. KNEIP (1999), On the estimation of jump points in smooth curves, *Annals of the Institute of Statistical Mathematics* **51**, 231–251.

HALL, P. and D. M. TITTERINGTON (1992), Edge preserving and peak preserving smoothing, *Technometrics* **34**, 429–440.

HINKLEY, D. V. (1969a), Inference in two phase regression, *Journal of the American Statistical Association* **66**, 736–743.

HINKLEY, D. V. (1969b), Inference about the intersection in two-phase regression, *Biometrika* **56**, 495–504.

JOSE, C. T. and B. ISMAIL (1999), Change points in nonparametric regression functions, *Communications in Statistics – Theory and Methods Series A* **28**, 1883–1902.

KHMALADZE, E., R. MNATSAKANOV and N. TORONJADZE (1997), *The change set problem. Part 1: Estimation of the change set and the rate of convergence*, School of Mathematics, The University of New South Wales, Sydney. Unpublished Manuscript.

KOCH, I. and A. POPE (1997), Asymptotic bias and variance of a kernel-based estimator for the location of a discontinuity, *Journal of Nonparametric Statistics* **8**, 45–64.

LIU, J., S. WU and J. V. ZIDEK (1997), On segmented multivariate regression, *Statistica Sinica* **7**, 497–525.

LOADER, C. R. (1996), Change point estimation using nonparametric regression, *Annals of Statistics* **24**, 1667–1678.

MÜLLER, H.-G. (1992), Change-points in nonparametric regression analysis. *Annals of Statistics* **20**, 737–761.

MÜLLER, H.-G. and K.-S. SONG (1997), Two-stage change-point estimators in smooth regression models, *Statistics and Probability Letters* **34**, 323–335.

QIU, P. and B. YANDELL (1998), A local polynomial jump-detection algorithm in nonparametric regression, *Technometrics* **20**, 141–152.

STUTE, W. and J.-L. WANG (1993), The strong law under random censorship, *Annals of Statistics* **21**, 1591–1607.

WU, J. S. and C. K. CHU (1993), Kernel type estimators of jump points and values of a regression function, *Annals of Statistics* **21**, 1545–1566.

Received: October 2000. Revised: May 2001.